Properties of Control Chart Zone Tests

By S. W. ROBERTS

(Manuscript received September 10, 1957)

This paper is concerned with the statistical properties of tests composed of the standard control chart test supplemented by one or more tests for runs of points in various zones into which the control chart is partitioned. The basic properties of the resultant tests, called zone tests, are illustrated graphically. A procedure for determining the properties of many zone tests of practical interest is described.

1. INTRODUCTION

1.1 *General*

In using an $\bar{X}$ control chart to maintain control of a process average, we periodically measure n units of the product and plot the average measurement $\bar{X}_n$ on the control chart in its chronological position. The control chart presents a pictorial summary of production history that is useful in: (a) detecting changes in the process average, and (b) providing clues to the causes of such changes. Various run tests have proved useful in application (b).[1] Most of the literature on run theory pertains to this application. There are tests for runs up and for runs up and down; there are tests for the number of runs and for the lengths of runs. The control chart is particularly suitable for run tests. We shall consider the use of a particular type of run test in application (a).

In application (a), as each point is plotted we decide whether or not to look for trouble (to take action to eliminate the cause of the change in the process average). Using the standard control chart test,[2, 3] we look for trouble if a point falls in a zone outside of two control limits symmetrically placed on either side of a line representing the nominal process average. The control limits, called the 3σ (3-sigma) limits, are placed at $\bar{X}_0' \pm 3(\sigma'/\sqrt{n})$, where $\bar{X}_0'$ and σ' are the nominal process average and standard deviation, respectively, and n is the sample size, or number of units of product measured for each point. We shall assume that $\bar{X}_0'$ and σ' are known, and that σ' remains fixed.

In using a statistical test to decide at each point whetber or not to look for trouble, we are subject to two types of errors:

(1) We make Type 1 errors when we decide to look for trouble when in fact none is present.

(2) We make Type 2 errors when we decide not to look for trouble when trouble is actually present.

Few Type 1 errors are made when the standard control chart test is used — an average of about one point in 370 falls outside of the 3σ limits when the process average is at its nominal level. Type 2 errors occur at consecutive points following a change until the test used indicates that a change has occurred. Small changes may result in long sequences of Type 2 errors because the probability of a point falling outside of the 3σ limits may be small, though larger than it was when the process average was at its nominal level. This definition of the two types of errors makes a sharp distinction between the presence and absence of trouble — a distinction more theoretical than practical — in order to simplify the exposition of the subject.

Experience indicates that, in general, the standard control chart test maintains an economic balance between the two types of errors in a wide range of industrial applications (Reference 2, pp. 276–7; Reference 3, p. 11). However, other tests may be more attractive economically in applications where early detection of relatively small changes is important. It has been suggested (Reference 4, p. 128) that supplementary run tests may prove useful in such applications. Various run tests are used in practice to supplement the standard control chart test,* but little has been published on the properties of the resultant tests,† though it is quite apparent that each additional supplementary run test employed decreases the number of Type 2 errors made and increases the number of Type 1 errors.

There are several alternative ways to reduce the number of Type 2 errors made; we can:

(1) Set the limit lines closer to the nominal process average $\bar{X}_0'$.

(2) Increase the sample size.[5,7]

(3) Replace the standard test with a single test for runs of points outside of appropriate limits.[6]

(4) Supplement the standard test with one or more run tests.

(5) Temporarily modify the sampling procedure — e.g., increase the sample size or frequency of sampling — whenever a point falls outside of "warning" limits but inside of the "action" limits (3σ limits).[4,7]

---

* See footnote, page 89.

† After the page proofs of this paper had been received, the author was advised of Reference 9, which deals primarily with the test $T_{12}(L_1, L_2)$.

(6) Use a control chart for a statistic other than $\bar{X}_n$; for example, plot points representing the moving average of k consecutive $\bar{X}_n$'s.
The improvements generally require extra information or more complicated tests, or they result in an increased frequency of Type 1 errors.

In this paper we study the properties of various run tests that either replace or supplement the standard control chart test in application (a). We limit our study to a particular type of run tests which we call "zone tests" because they test for runs of points in various zones into which the control chart is partitioned. For example, we study such tests as $T_{12'}$ (3, 2),† which calls for action if a single point falls outside of the 3σ limits or if two of three consecutive points fall outside of a 2σ limit line. We limit our studies to tests used on charts of the statistic $\bar{X}_n$; zone tests can be useful on other charts, but their properties depend on the properties of the particular statistic plotted. Our results apply for any sample size and frequency of sampling.

We use $T_k(L_k)$ to denote a test for k consecutive points outside of one of the pair of limit lines at $\bar{X}_0' \pm L_k(\sigma'/\sqrt{n})$, and $T_{k'}(L_k)$ to denote the test for k out of $k + 1$ consecutive points outside of the limit lines. If we combine two tests, we let $T_{k_1k_2}(L_{k_1}, L_{k_2})$ denote the combined test that calls for action on the occurrence of *either* type of run; k_1 and k_2 are integers less than nine, either primed or unprimed.

For simplicity of notation we may eliminate the brackets on the test notation if the subscripts provide sufficient information. For this purpose, we adopt standard limits for certain runs. Thus we may use T_1 rather than $T_1(3)$ to denote the standard control chart test. Also, we use the 2σ limits, the 1σ limits, and $\bar{X}_0'$ itself as standard for runs of lengths 2, 4, and 8, respectively. Thus $T_{12'}$ means $T_{12'}(3, 2)$, and T_8 means $T_8(0)$. We use an asterisk to denote one-sided tests — those with limit lines on only one side of $\bar{X}_0'$. Test T_1^* has a single limit line, at $\bar{X}_0' + 3(\sigma'/\sqrt{n})$.

1.2 *Process Model*

We use a process model in which the process average is $\bar{X}' = \bar{X}_0' + \Delta$, where Δ is subject to change. A picture showing how Δ changes with time would show a series of rectangular pulses (positive or negative) of various heights, separated by periods with $\Delta = 0$. The beginning of a pulse corresponds to the occurrence of an assignable cause of variation, and the height of the pulse is a function of the particular cause. The pulse ending corresponds to the elimination of the trouble. The distribution of

† Read subscript as 1, 2'.

the lengths of the pulses depends on the test we use to detect changes; the test should be designed to keep the lengths reasonably short.

The sample average, $\bar{X}_n$, is assumed to have a normal distribution defined by its expected value $\bar{X}'$ and standard deviation $\sigma'/\sqrt{n}$.

Whenever $\Delta = 0$, the process average is at its nominal level, and we say the process is in State 1. Whenever $\Delta \neq 0$, there is trouble present, and we say the process is in State 2. We assume that no additional changes occur while the process remains in State 2.

At each point we look for certain runs that rarely occur in State 1. In the absence of such runs there is no indication that the process is not in State 1, and accordingly we do not look for trouble. We do not attempt to define the probability that the process is in State 1 at any point. In this model, we stop the process to look for trouble on the *first* occurrence of a run for which we are testing. When the process starts again it is assumed to be in State 1; consequently, the testing procedure ignores previous points.

Relatively straight-forward mathematics can be used to describe the properties of certain tests acting within the framework of this process model. Alternative, and perhaps more realistic, assumptions can easily lead to much more complicated problems of description. In many cases the results obtained here can be used to describe qualitatively the properties of tests applied to more complex processes.

1.3 *Measuring the Two Types of Decision Errors*

As each point is plotted on the control chart we decide either that the process is in State 1 — in which case we leave it alone — or that it is in State 2 — in which case we look for trouble. We make a Type 1 error when we say that the process is in State 2 when actually it is in State 1; Type 1 errors initiate needless action. We make a Type 2 error when we say that the process is in State 1 when actually it is in State 2; Type 2 errors fail to initiate needed action. We generally make a series of consecutive errors of Type 2 before detecting the change in state.

Let the random variable y denote the number of points plotted while the process remains in State 2. Then $y - 1$ consecutive errors of Type 2 are made. Let $E(y)$ denote the expected, or average, value of y; then $E(y - 1)$ is the average length of a series of Type 2 errors.

$E(y)$ depends on Δ, the amount by which the process average changes; we sometimes note this dependence by writing $E(y;\, \Delta)$. $E(y;\, \Delta)$ is a monotonically decreasing function of the magnitude of Δ; that is, the larger the change, the smaller is $E(y)$. In other words, tests are more sensitive to large changes than to small changes.

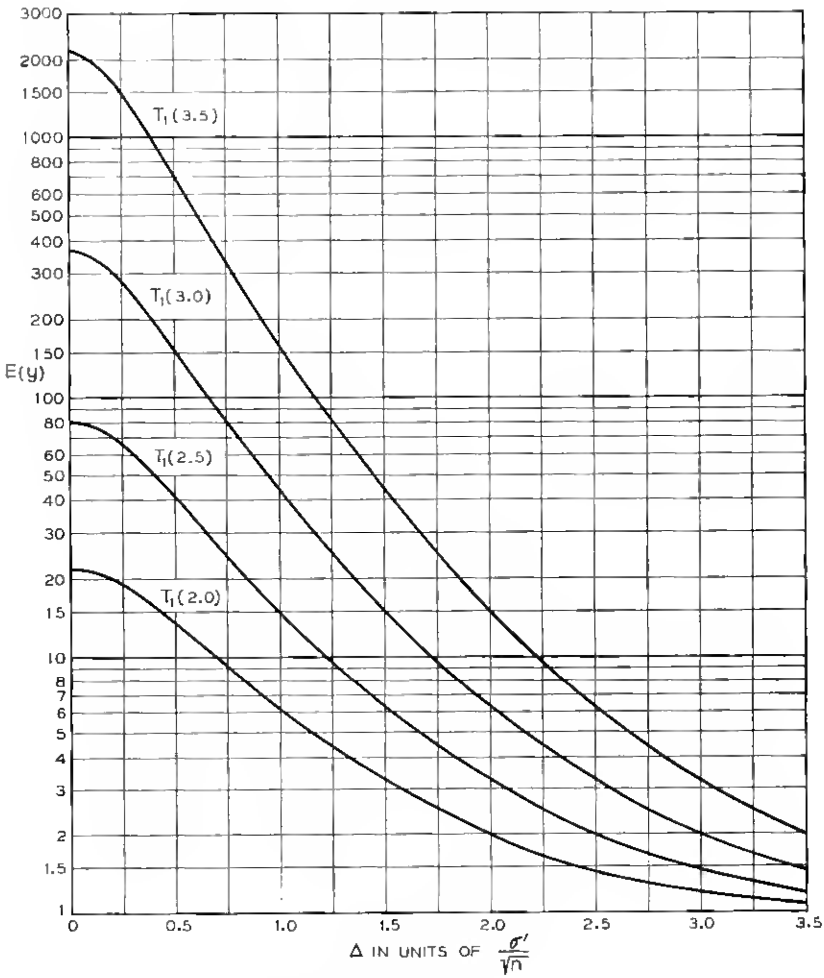

Fig. 1 — $E(y)$ versus Δ for $T_1(L_1)$ for various limits.

Fig. 1 shows curves of $E(y)$ versus Δ for $T_1(L_1)$ for $L_1 = 2, 2.5, 3$, and 3.5. Note on the curve for $T_1(3)$, for example, that $E(y) = 15$ at $\Delta = 1.5\ (\sigma'/\sqrt{n})$; this means that following a change of this magnitude, an average of 15 points are plotted before a point falls outside of a 3σ limit. Note that as Δ approaches zero, $E(y)$ approaches 370, which corresponds to the average number of points between consecutive Type 1 errors while the process remains in State 1.

In Fig. 1 and later figures the abscissa is Δ, and it is measured in units of $\sigma'/\sqrt{n}$, which is the standard deviation of $\bar{X}_n$. The particular ab-

scissa that applies to a change of a given physical magnitude is proportional to $\sqrt{n}$; for example, if n is doubled in the above example where $\Delta = 1.5(\sigma'/\sqrt{n})$, then the appropriate abscissa on Fig. 1 increases from 1.5 units, with ordinate $E(y) = 15$ on curve T_1, to $1.5\sqrt{2} = 2.121$ units, with $E(y) = 5.4$. The positions of the curves relative to one another are independent of n.

If the process were to remain in State 1 indefinitely, $E(y; 0)$ would represent the average number of points between consecutive Type 1 errors, and $1/[E(y; 0)]$ would be the asymptotic probability of a Type 1 error. In comparing tests with respect to Type 1 errors, we compare their values of $E(y; 0)$.

In comparing tests with respect to Type 2 errors, we compare their values of $E(y)$, or $E(y - 1)$, for various non-zero values of Δ.

1.4 *Comparing the Statistical Properties of Various Zone Tests*

We are primarily interested in the distribution of y. The distribution of y for all of the zone tests we consider can be adequately summarized by one parameter — its average value $E(y)$ (see Section 3.1). Therefore, in comparing the statistical properties of various tests, we compare their curves of $E(y)$ versus Δ. From such curves we can determine the asymptotic probability of Type 1 errors, $1/[E(y; 0)]$, and the average number of consecutive Type 2 errors, $E(y - 1; \Delta)$, for any Δ different from zero.

Figure 1 illustrates how the properties of zone tests can be changed by changing the limit lines. By changing the limit lines of $T_1(L_1)$ from $L_1 = 3$ to $L_1 = 2$, we reduce $E(y)$ for all values of Δ: when $\Delta > 0$, this means that the Type 2 errors are reduced; when $\Delta = 0$, this means that Type 1 errors are increased.

A choice between two tests should be based partially on the relative values of the two types of decision errors. We can fix the Type 1 errors at any desired level by an appropriate setting of the zone limits; then the Type 2 errors alone serve as a basis of comparison.

II. SUMMARY OF RESULTS

Section 4 shows how to determine the distribution of y, and in particular its average value $E(y)$, for one-sided tests $T_k^*(L_k)$ and $T_{k'}^*(L_k)$, for any k. Simple substitutions into equations for the above one-sided tests allow us to determine the properties of any test of type $T_{1k}^*(L_1, L_k)$ or $T_{1k'}^*(L_1, L_k)$. We then determine the properties of two-sided tests from the properties of the corresponding one-sided tests.

We show that the average values of y in any two separate tests provide upper and lower bounds to the average value of y in their combined test. Thus with subscript t_1 denoting test T_{t_1}, t_2 denoting T_{t_2}, and t_1t_2 de-

noting the test $T_{t_1 t_2}$ combining T_{t_1} and T_{t_2}, we have upper bounds

$$E_{t_1 t_2}(y; \Delta) \leqq E_{t_1}(y; \Delta), \qquad E_{t_1 t_2}(y; \Delta) \leqq E_{t_2}(y; \Delta), \tag{1}$$

and a lower bound

$$\frac{1}{E_{t_1 t_2}(y; \Delta)} \leqq \frac{1}{E_{t_1}(y; \Delta)} + \frac{1}{E_{t_2}(y; \Delta)}. \tag{2}$$

An application of (2) to the determination of the properties of two-sided tests in terms of the properties of their component one-sided tests yields

$$\frac{1}{E(y; \Delta)} \cong \frac{1}{E^*(y; \Delta)} + \frac{1}{E^*(y; -\Delta)}, \tag{3}$$

where the asterisks denote one-sided test results.

We can determine the properties of the following tests: $T_k(L_k)$, $T_{k'}(L_k)$, $T_{1k}(L_1, L_k)$ and $T_{1k'}(L_1, L_k)$, for any k. With $L_1 = 3$, the last two types of tests supplement the standard control chart test $T_1(3)$ with one other zone test.

Equations (1) support the logical conclusion that the more criteria we have to indicate the presence of trouble, the more quickly we will look for trouble when it is present as well as when it is not present. Thus, in supplementing the standard control chart test with other tests, we decrease the Type 2 errors at the expense of more frequent Type 1 errors. The question of how far to go in supplementing the standard control chart test must be answered in light of the relative importance of the two types of errors in the particular application considered.

Section III presents a series of charts to show the properties of several particular tests, including T_1, T_{12}, $T_{12'}$, T_{18}, and $T_{12'4'8}$. The last test* illustrates the effect of supplementing T_1 with more than one additional test; its properties were determined through the use of Monte Carlo techniques. We also show $E(y)$ versus Δ for several tests when their zone limits are translated away from the center line so that their Type 1 errors are comparable to those of T_1. It is through such translations of zone limits that we can offset the undesirable effect on Type 1 errors that occurs when we add new tests to our testing procedure.

III. CHARTS SHOWING PROPERTIES OF VARIOUS ZONE TESTS

3.1 *Distribution Function of y*

The cumulative distribution function of the random variable y, $Q_j = \text{Prob}\ (y > j)$, is shown in Figs. 2 and 3 for various zone tests.

* This test is similar to one that has been used by the Western Electric Company in its quality control training program; somewhat different criteria for taking action are used and therefore the statistical properties differ.

The curves are applicable only at integral values of j. If $y > j$, there have been no indications of a changed process average in the first j points following the change from $\bar{X}_0'$ to $\bar{X}_0' + \Delta$.

Fig. 2 shows curves for T_1, $T_{2'}$, $T_{4'}$, and T_8 for $\Delta = 0, \sigma'/\sqrt{n}$, and $2(\sigma'/\sqrt{n})$. Fig. 3 compares T_1, $T_{2'}$, $T_{12'}$, and $T_{12'4'8}$ for $\Delta = \sigma'/\sqrt{n}$; it illustrates the effect of additional tests on the distribution of y.

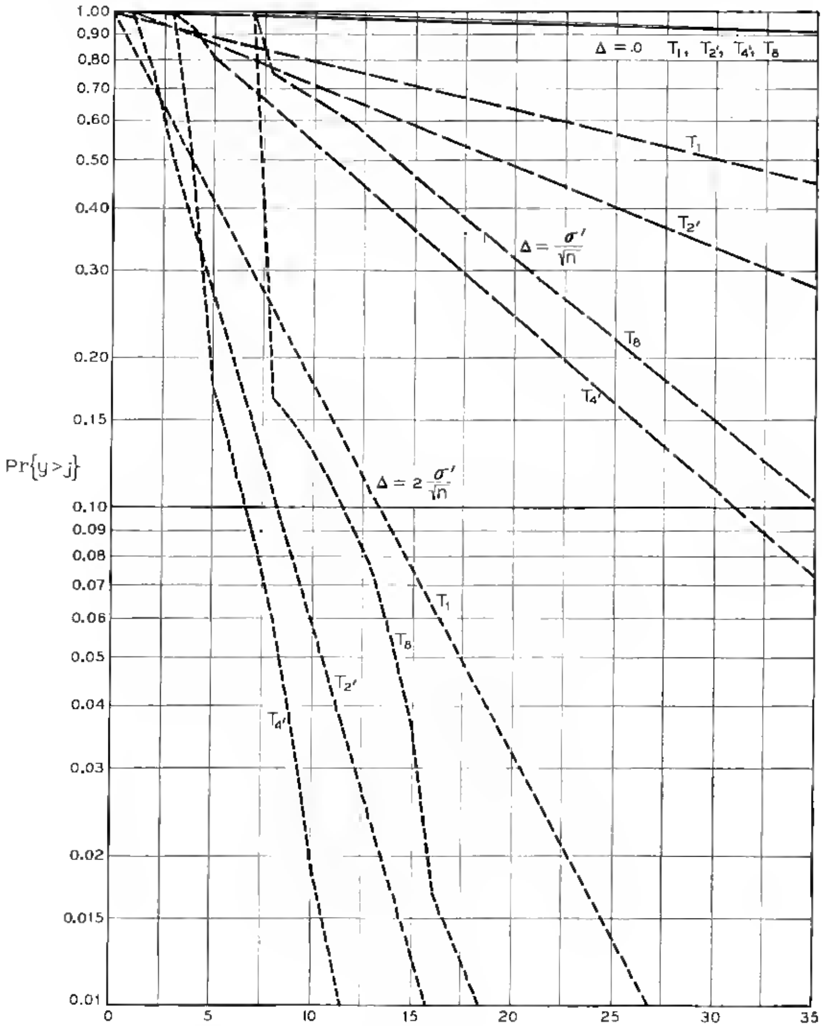

Fig. 2 — Cumulative distribution of y for T_1, $T_{2'}$, $T_{4'}$, and T_8 for $\Delta = 0, \sigma'/\sqrt{n}$, and $2(\sigma'/\sqrt{n})$.

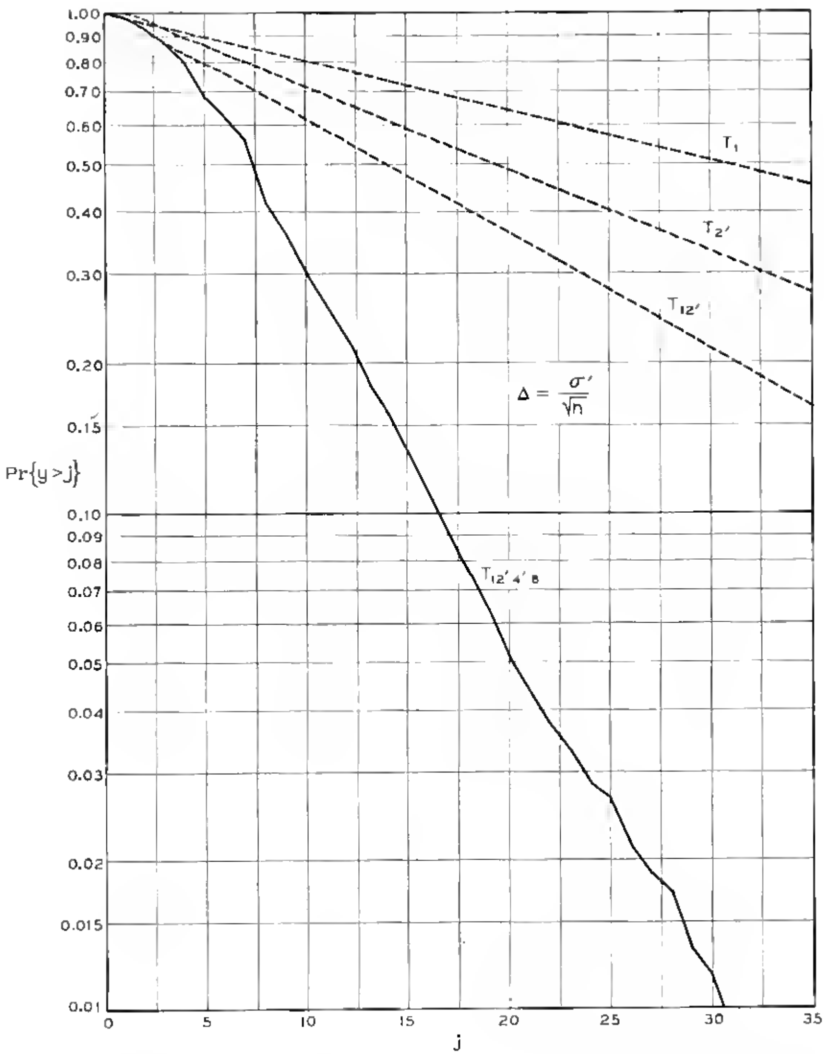

Fig. 3 — Cumulative distribution of y for T_1, $T_{2'}$, $T_{12'}$ and $T_{12'4'8}$ for $\Delta = \sigma'/\sqrt{n}$.

The curves of Figs. 2 and 3, plotted on semilogarithmic paper, can be approximated for practical purposes by straight lines. Thus, the distributions are approximately geometric, or discrete exponential, distributions that can be described by a single parameter $E(y)$ and an initial value. It is for this reason that $E(y)$ adequately summarizes their statistical properties.

Figs. 2 and 3 illustrate the fact that single tests for long runs, such as T_8, do not become fully effective immediately following a change.

3.2 *Curves of $E(y)$ Versus Δ for Tests with Standard Zone Limits*

Figs. 4, 5, and 6 illustrate typical curves of $E(y)$ versus Δ. Fig. 4 shows curves for T_1, $T_{2'}$, $T_{4'}$, and T_8. Fig. 5 shows the effect of broadening the criteria for looking for trouble — T_2 calls for action only if two consecutive points fall outside of a 2σ limit, whereas $T_{2'}$ calls for

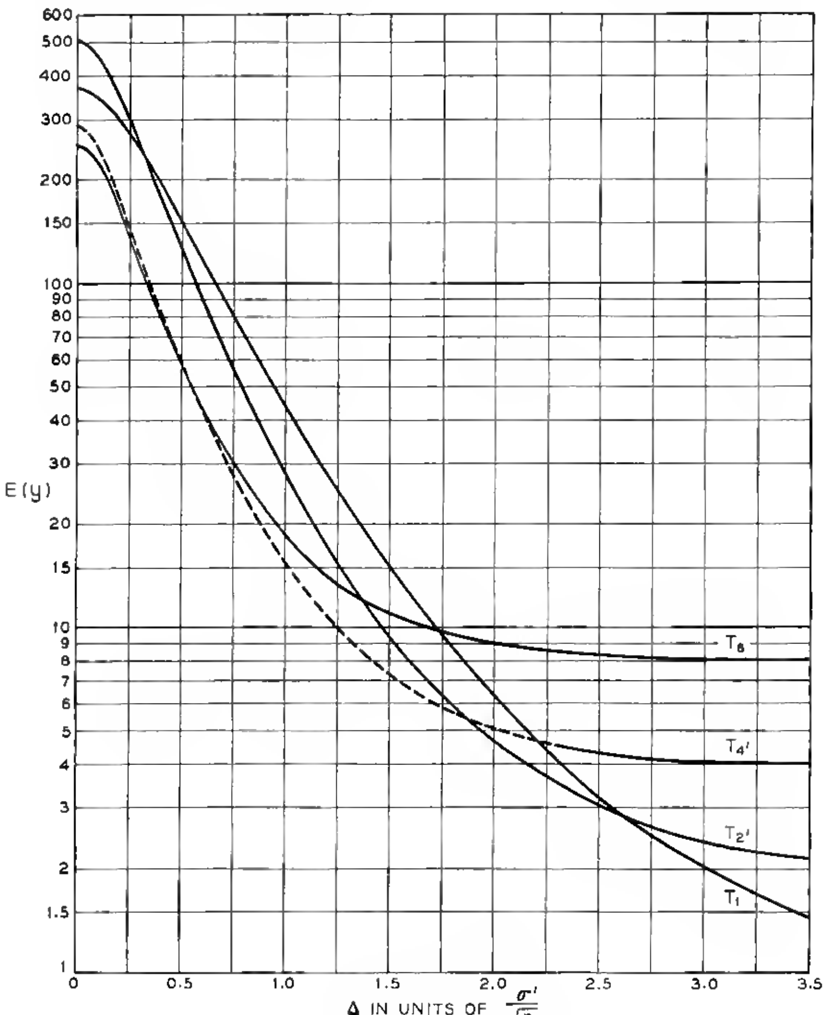

Fig. 4 — $E(y)$ versus Δ for T_1, $T_{2'}$, $T_{4'}$, and T_8.

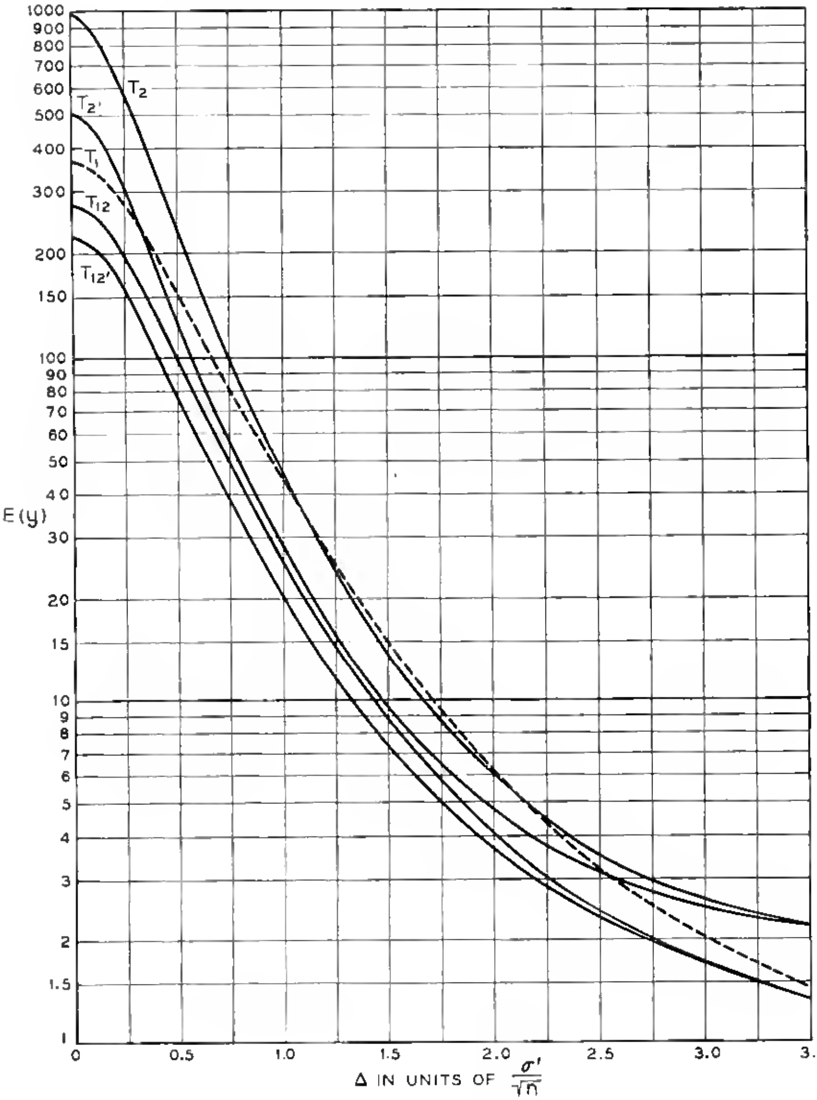

Fig. 5 — $E(y)$ versus Δ for T_1, T_2, $T_{2'}$, T_{12}, and $T_{12'}$.

action whenever T_2 does and also whenever two points falling outside a 2σ limit are separated by a single point not falling outside of the 2σ limit. $E(y)$ is less for $T_{2'}$ than for T_2 for all values of Δ; this difference is reflected in the curves for T_{12} and $T_{12'}$, which supplement T_1 with T_2 and $T_{2'}$, respectively.

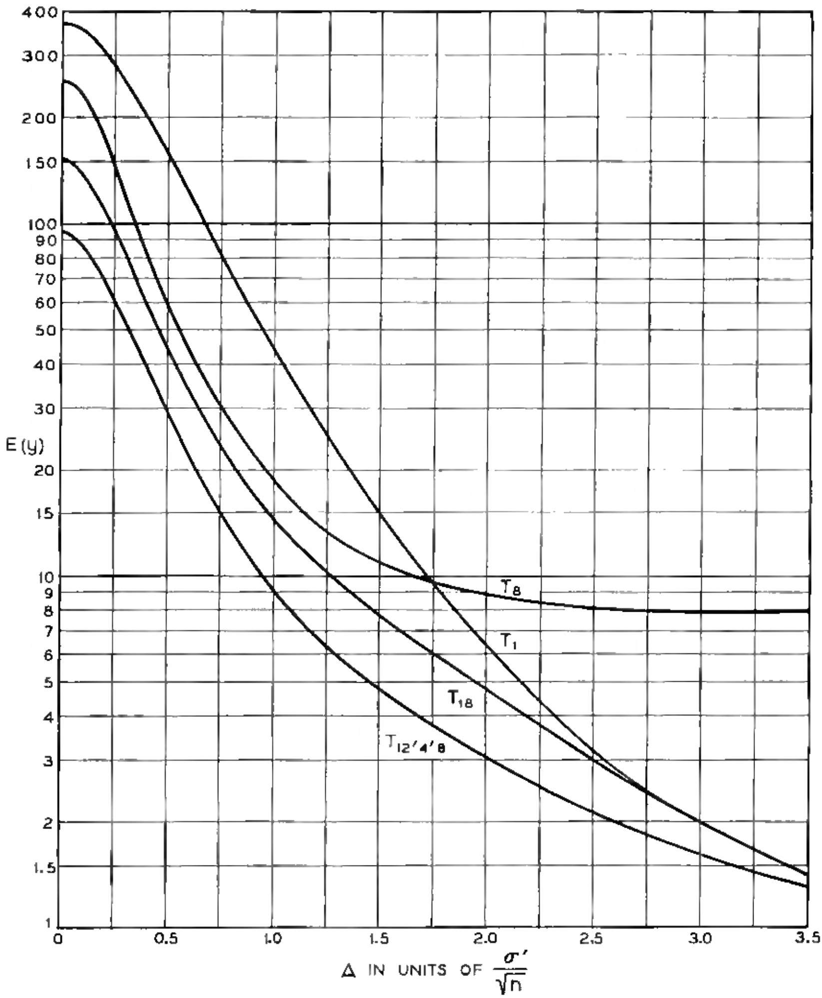

Fig. 6 — $E(y)$ versus Δ for T_1, T_8, T_{18}, and $T_{12'4'8}$.

Fig. 6 illustrates the effect of supplementing T_1 first with T_8 and then with T_8, $T_{2'}$, and $T_{4'}$. Notice how the Type 1 errors become more frequent as Type 2 errors decrease.

3.3 *Curves of $E(y)$ Versus Δ with Limits Set for a Selected Probability of Type 1 Errors*

Fig. 7 shows curves of $E(y)$ versus Δ for tests for k ($k = 1, 2, 3, 4, 6, 8$) consecutive points outside of limits that are set for each k so that the

probability of a Type 1 error is comparable to that of T_1. Tests for long runs clearly are most effective against small process changes, while T_1 itself is most effective against large process changes.

Fig. 8 shows curves for T_1, $T_8(0.065)$, and $T_{18}(3.19, 0.19)$. The last test is composed of the first two tests with all zone limits translated away from $\bar{X}_0'$. Notice that T_1 and $T_8(0.065)$ taken individually are more effective than $T_{18}(3.19, 0.19)$ in certain ranges of Δ. Fig. 9 illustrates

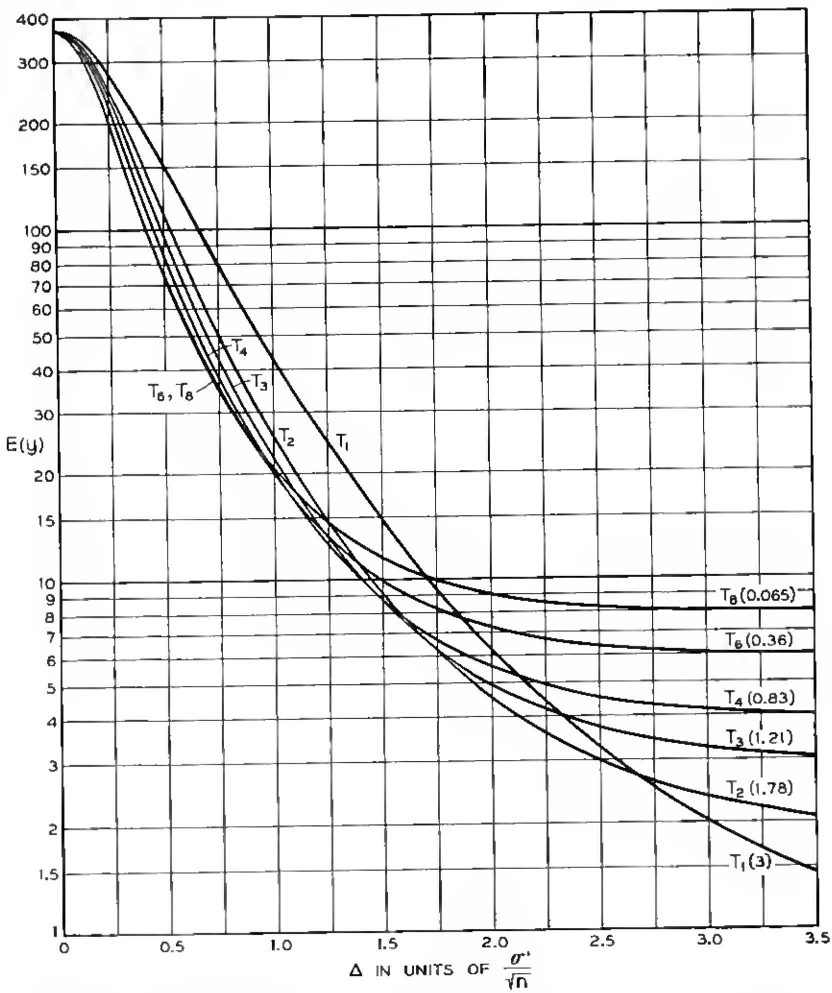

Fig. 7 — $E(y)$ versus Δ for $T_k(L_k)$ for $k = 1, 2, 3, 4, 6$, and 8 with limits set for the same probabilities of Type 1 errors.

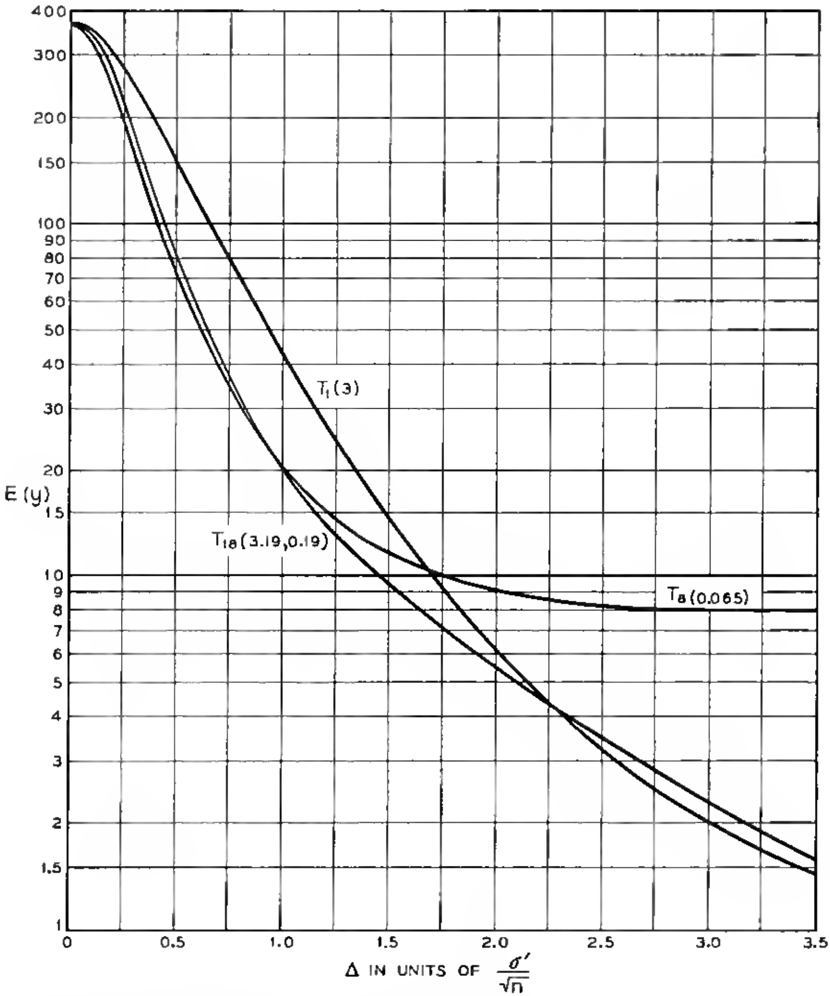

Fig. 8 — $E(y)$ versus Δ for T_1, $T_8(L_8)$, and $T_{18}(L_1, L_8)$ with limits set for the same probabilities of Type 1 errors.

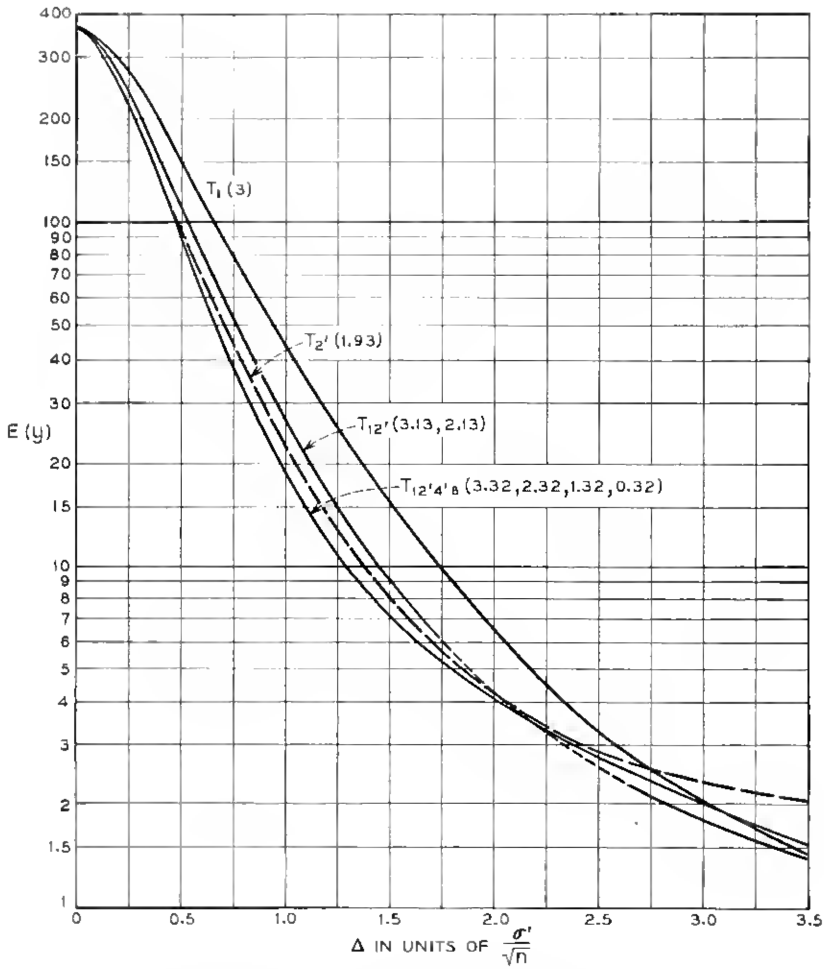

Fig. 9 — $E(y)$ versus Δ for T_1, $T_{2'}(L_{2'})$, $T_{12'}(L_1, L_{2'})$, and $T_{12'4'8}(L_1, L_{2'}, L_{4'}, L_8)$ with limits set for the same probabilities of Type 1 errors.

the same general ideas as Fig. 8, with the addition of $T_{12'4'8}$ with its zone limits translated away from $\bar{X}_0'$.

Because logarithmic scales are used for $E(y)$, the differences $E_1(y) - E_t(y)$ between curves for T_1 and other tests are distorted; Fig. 10 shows the difference on an arithmetic scale for two of the curves of Fig. 9.

Fig. 11 supports the theory that $T_{k'}(L_{k'})$ is slightly more sensitive to small changes than $T_k(L_k)$ when the limits are set so that the two tests have the same probabilities of Type 1 errors. Further graphical

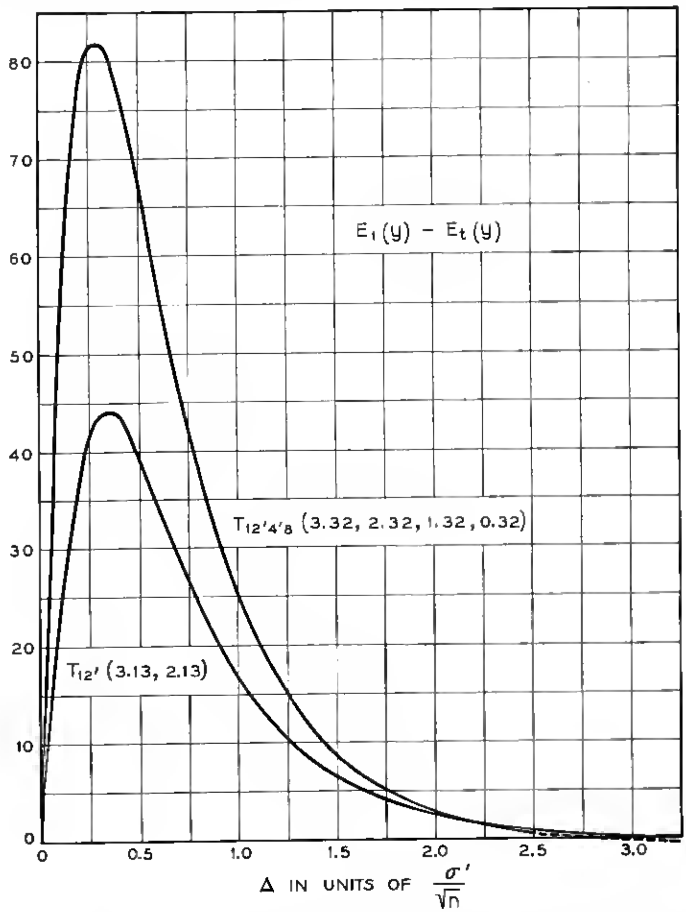

Fig. 10 — The difference between ordinates of curves of Fig. 9 shown on an arithmetic scale.

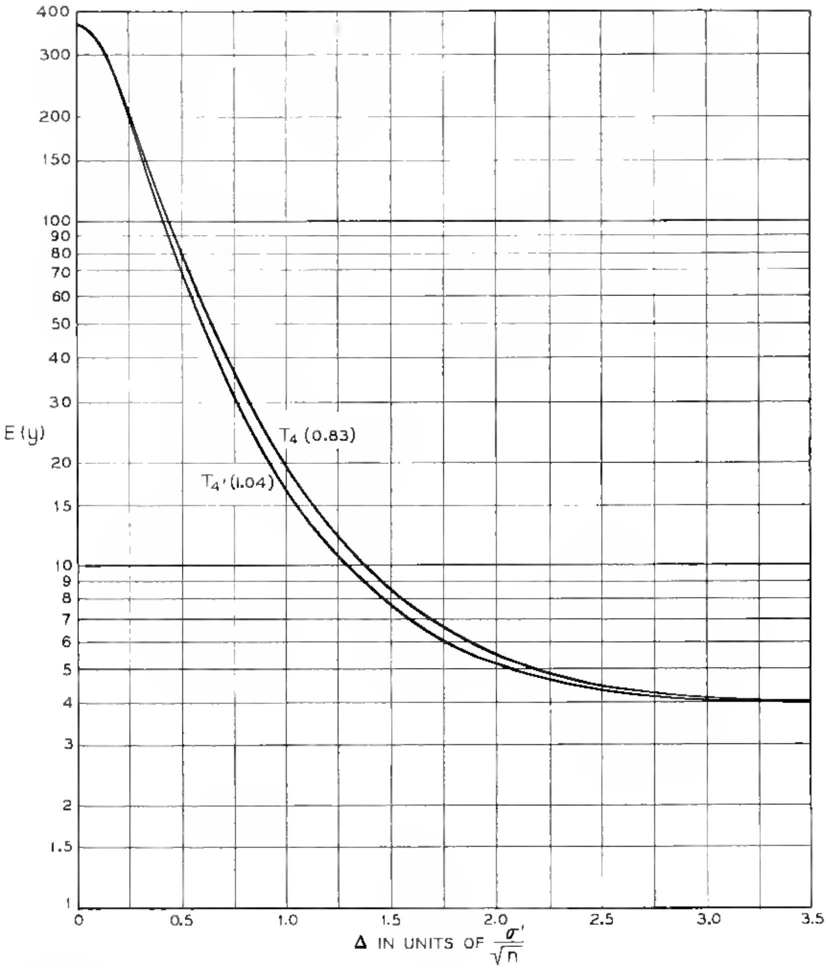

Fig. 11 — $E(y)$ versus Δ for $T_4(L_4)$ and $T_{4'}(L_{4'})$ with limits set for the same probabilities of Type 1 errors.

support is given by curves for $T_{2'}(1.93)$ of Fig. 9 and $T_2(1.78)$ of Fig. 7. No analytical proof has been developed.

IV. DETERMINING THE STATISTICAL PROPERTIES OF ZONE TESTS

4.1 *General Procedure*

With the control chart partitioned into mutually exclusive zones A, B, C, D, $\cdots$, R, we represent a sequence of points falling consecutively

into zones B, C, D, and B, for example, by the sequence $bcdb$. The lower case letters such as b serve a dual purpose — they denote the fact that a point falls into a particular zone, and they denote the probability of that particular event, or outcome. For example, the probability of a particular sequence $bcdbcdb$ is $b^3c^2d^2$. Where there is danger of confusion, we may denote outcome b by ϵ_b and its probability by p_b. A sequence $bcdb$ is considered to represent the outcome of a sequence of independent trials, each of which has fixed probabilities of outcomes $a,b,c, \cdots, r$.

Since the control chart points represent an average measurement $\bar{X}_n$ that has a normal distribution with average $\bar{X}_0' + \Delta$ and standard deviation $\sigma'/\sqrt{n}$, we use normal probability tables to determine the probability b, which remains constant from point to point as long as the process remains in a given state. If $\Phi(x)$ is the area under the normal curve above x, and if zone B is between limit lines at $\bar{X}_0' + L_2(\sigma'/\sqrt{n})$ and $\bar{X}_0' + L_1(\sigma'/\sqrt{n})$, where $L_2 \leqq L_1$, then probability

$$b = \Phi\left(L_2 - \frac{\Delta\sqrt{n}}{\sigma'}\right) - \Phi\left(L_1 - \frac{\Delta\sqrt{n}}{\sigma'}\right) .$$

When the process changes from State 1 to State 2, the probabilities of points falling into the various zones change. At the first point in State 2, zone tests see one point from State 2 preceded by a sequence of points from State 1; at each subsequent point in State 2 a single point from State 1 is dropped from consideration, until at last all points considered are from State 2. The zone tests are such that the probability of a point from State 2 falling into a critical zone is greater than the probability of a point from State 1 falling into the same zone. Consequently, the probability of the occurrence of a run of points in a critical zone is greatest if all of the points are from State 2. For simplicity and clarity we neglect points from State 1 while considering the results of testing points from State 2. This means that T_8, for example, does not become effective until the eighth point in State 2 appears. This simplifying assumption will affect the results little; its effect can be eliminated by calculating the probability of detecting the change in the first few points and adjusting our results. As an illustration, $T_8(0.065)$ of Fig. 7 should approach 7.1, rather than 8, as Δ approaches infinity.

If a control chart is partitioned into three mutually exclusive zones A, B, and C, outcomes a, b, and c are associated with the events that points fall in the respective zones, and probabilities a, b, and c $(a + b + c = 1)$ are the corresponding probabilities of the events, or outcomes. The possible outcome of the first j trials, or points, can be enumerated by the ordered expansion of the multinomial $(a + b + c)^j$.

For example, with $j = 2$, we have:

$$(a + b + c)^2 = aa + ab + ac + ba + bb + bc + ca + cb + cc.$$

The probabilities of the various sequences occurring are obtained simply by multiplying the individual terms; for example, sequence aa has probability a^2. The probability of a particular event such as the event that either a or b occurs at least once in the first two trials is determined by selecting those sequences in which this event occurs and cumulating their probabilities; in this case it is $a^2 + b^2 + 2ab + 2ac + 2bc$.

If we wished to determine the probability Q_j of no occurrences in the first j trials of an event ϵ (a run of eight consecutive points in zone A, for example), we could enumerate all of the 3^j possible outcomes, pick out those we were interested in, and determine their probabilities. This procedure becomes very tedious as j increases, and we soon look for shortcuts. We attempt to find a recursion equation defining Q_j in terms of a limited number of terms Q_{j-1}, Q_{j-2}, etc. If we can find such an equation, we need to enumerate all pertinent outcomes only to the point where the equation becomes effective.

A recursion equation for Q_j, together with a set of initial conditions, leads to a generating function $Q(s)$ whose power series expansion exhibits Q_j as the coefficient of s^j:

$$Q(s) = 1 + Q_1 s + Q_2 s^2 + \cdots + Q_j s^j + \cdots = \sum_{j=0}^{\infty} Q_j s^j. \tag{4}$$

The generating function is useful in obtaining moments of the distribution of y. In particular, we obtain $E(y)$ by setting $s = 1$ in the equation for $Q(s)$: $E(y) = Q(1)$.

The simplest zone tests are those in which a point is classified in one of two categories; it represents either event ϵ_p with probability p or event ϵ_q with probability $q = 1 - p$. We arbitrarily call ϵ_p a success and ϵ_q a failure.* We call a test for success runs a simple run test. A compound run test is composed of more than one simple run test; for example, a test for a run of two consecutive points above the $+2\sigma$ limit is a simple run test, but a test for a run of two consecutive points above the $+2\sigma$ limit *or* below the -2σ limit is a compound run test composed of two simple run tests. A simple run test classifies points in two ways; a compound run test classifies points in more than two ways.

The test for a run of two consecutive points above the $+2\sigma$ limit is a one-sided zone test; the test for a run of two consecutive points above

* This terminology may seem incongruous, since we hope for events ϵ_q, which we term failures. Alternatively, we could change the definition, and say that we test for failure runs, but this conflicts with standard terminology.

the $+2\sigma$ limit or below the -2σ limit is a two-sided zone test. We derive the properties of two-sided tests from those of one-sided tests.

In Sections 4.2 and 4.3 we present recursion equations and generating functions for Q_j, the probability that $y > j$, for the following simple run tests:

(1) k consecutive successes $\quad k = 1, 2, 3, 4, \cdots,$

(2) k successes in $k + 1$ (or k) consecutive trials $\quad k = 2, 3,$ and 4.

In addition, we describe a procedure for extending k in (2) to any value. Equations for $E(y)$ are also presented. The results apply to one-sided zone tests.

Section 4.4 describes a procedure for determining the properties of two-sided zone tests from the properties of one-sided zone tests.

Section 4.5 presents a procedure for determining the properties of any run test combined with a test for a single point in a critical zone. Simple substitutions into the equations for a particular one-sided zone test lead to a description of the properties of that test in combination with the standard control chart test T_1^*.

Section 4.6 develops upper and lower bounds to $E(y)$. Section 4.7 shows how to determine easily the properties of some tests whose zone limits are non-standard. Section 4.8 discusses the use of Monte Carlo techniques for determining the properties of tests more complex than those considered here.

4.2 *The First Occurrence of k Consecutive Successes*

We separate those sequences of outcomes having no occurrences of k consecutive successes in the first j trials (that is, $y > j$) into mutually exclusive categories according to whether the last failure occurred on trial $j, j - 1, j - 2, \cdots$ or $j - k + 1$. With Q_j denoting the probability that $y > j$, we let $Q_{j,i}$ denote the probability that $y > j$ and that trial $j - i$ resulted in a failure and the succeeding i trials resulted in successes. Then, since i can be no greater than $k - 1$, we have the equation:

$$Q_j = Q_{j,0} + Q_{j,1} + Q_{j,2} + \cdots + Q_{j,k-1}. \tag{5}$$

We enumerate the possible results:

Sequence Endings	Probabilities of Occurrence
$\cdots\cdots\cdots q$	$Q_{j,0} = q(Q_{j-1,0} + Q_{j-1,1} + \cdots + Q_{j-1,k-1})$
$\cdots\cdots\cdots qp$	$Q_{j,1} = pq(Q_{j-2,0} + Q_{j-2,1} + \cdots + Q_{j-2,k-1})$
$\cdots\cdots qpp$	$Q_{j,2} = p^2q(Q_{j-3,0} + Q_{j-3,1} + \cdots + Q_{j-3,k-1})$
$\cdots\cdots qppp$	$Q_{j,3} = p^3q(Q_{j-4,0} + Q_{j-4,1} + \cdots + Q_{j-4,k-1})$
$\vdots$	
$\underbrace{qpp \cdots p}_{k-1\ p\text{'s}}$	$Q_{j,k-1} = p^{k-1}q(Q_{j-k,0} + Q_{j-k,1} + \cdots + Q_{j-k,k-1})$

The equations on the right reduce to $Q_{j,i} = p^i q Q_{j-i-1}$. We obtain the desired recursion equation by summing over all values of i,

$$Q_j = qQ_{j-1} + pqQ_{j-2} + p^2qQ_{j-3} + \cdots + p^{k-1}qQ_{j-k}. \tag{6}$$

We can use (6) to calculate Q_j for $j \geqq k$, noting that $Q_j = 1$ for $j < k$.

We obtain the generating function of Q_j from (6),

$$Q(s) = \frac{1 - p^k s^k}{1 - s + qp^k s^{k+1}}. \tag{7}$$

Then $E(y)$ is obtained by setting $s = 1$ in (7),

$$E(y) = \frac{1 - p^k}{qp^k}. \tag{8}$$

These results are well known.[8]

4.3 *The First Occurrence of k Successes in $k + 1$ Consecutive Trials*

As in the preceding section, we separate those sequences having no occurrence of the event in question — in this case k successes in $k + 1$ consecutive trials — into mutually exclusive categories according to whether the last failure occurred on trial $j, j - 1, j - 2, \cdots,$ or $j - k + 1$. In the current problem, however, we are also interested in the location of the next-to-the-last failure since if the event in question has not occurred there must be at least two failures in the preceding $k + 1$ trials. If the last failure was on trial $j - (k - 2)$, for example, there must be at least one other failure in the preceding two trials. Here an enumeration of possible results yields:

Sequence Endings — Probabilities of Occurrence

$$\begin{array}{llll}
\cdots\cdots q & Q_{j,0} = q(Q_{j-1,0} & + Q_{j-1,1} + \cdots + Q_{j-1,k-2} + Q_{j-1,k-1}) & (9.0) \\
\cdots\cdots qp & Q_{j,1} = pq(Q_{j-2,0} & + Q_{j-2,1} + \cdots + Q_{j-2,k-2}) & (9.1) \\
\cdots\cdots qpp & Q_{j,2} = p^2q(Q_{j-3,0} & + \cdots + Q_{j-3,k-3}) & (9.2) \\
\vdots & \vdots & & \\
qp \cdots ppp & Q_{j,k-2} = p^{k-2}q(Q_{j-(k-1),0} + Q_{j-(k-1),1}) & & (9.(k-2)) \\
q\underbrace{pp \cdots ppp}_{k-1\ p\text{'s}} & Q_{j,k-1} = p^{k-1}q(Q_{j-k,0}). & & (9.(k-1))
\end{array}$$

Each equation in (9) has one term less than the equation immediately above it. We adopt a standard procedure for deriving a recursion equation for Q_j from equations (9). First we find from (9.0) that:

$$Q_{j,0} = qQ_{j-1}. \tag{10}$$

Then we substitute (10), with j reduced by k, into (9.$(k - 1)$):

$$Q_{j,k-1} = p^{k-1}q^2Q_{j-k-1}. \tag{11}$$

Next we translate the final term on the right-hand side of (9.0) to the left-hand side, and substitute (10) and (11), the latter with j reduced by one. Then, if we multiply through the new (9.0) by p and reduce j by one, its right-hand side is identical to that of (9.1). Then we have

$$Q_{j,1} = pqQ_{j-2} - p^k q^3 Q_{j-k-2} . \tag{12}$$

We substitute (10) and (12), with j reduced by $(k - 1)$, into $(9.(k - 2))$ to obtain

$$Q_{j,k-2} = p^{k-2}q^2Q_{j-k} + p^{k-1}q^2Q_{j-k-1} - p^{2k-2}q^4Q_{j-2k-2} . \tag{13}$$

We proceed step by step, taking equations from the top and then from the bottom, to find equations for the $Q_{j,i}$'s in terms of Q_j's. Then we add all of the equations together to obtain the recursion equation for Q_j, which will depend on some of the $k(k + 1)/2$ immediately preceding Q_j's. The recursion equation is used with $k(k + 1)/2$ initial Q_j's to derive the generating function $Q(s)$.

4.31 *The First Occurrence of Two Successes in Three Consecutive Trials*

As in (9), we have:

Sequence Endings	Probabilities of Occurrence	
$\cdot q$	$Q_{j,0} = q(Q_{j-1,0} + Q_{j-1,1})$	(14.0)
qp	$Q_{j,1} = pq(Q_{j-2,0})$.	(14.1)

Then $Q_{j,0} = qQ_{j-1}$, $Q_{j,1} = pq^2Q_{j-3}$, and the recursion equation is

$$Q_j = qQ_{j-1} + pq^2Q_{j-3}, \qquad j > 2. \tag{15}$$

With (15) and the initial conditions $Q_0 = Q_1 = 1$ and $Q_2 = 1 - p^2$, we derive the generating function for Q_j:

$$Q(s) = \frac{1 + ps + pqs^2}{1 - qs - pq^2s^3}. \tag{16}$$

$E(y)$ is obtained by setting $s = 1$ in (16);

$$E(y) = \frac{1 + p + pq}{p^2(1 + q)}. \tag{17}$$

4.32 *The First Occurrence of Three Successes in Four Consecutive Trials*

The initial conditions are:

$$Q_0 = Q_1 = Q_2 = 1,$$
$$Q_3 = 1 - p^3,$$
$$Q_4 = 1 - p^3 - 3p^3q,$$
$$Q_5 = 1 - p^3 - 3p^3q - 3p^3q^2.$$

For $j > 5$ we follow the standard procedure to find a recursion equation for Q_j in terms of the $3\cdot4/2 = 6$ preceding Q_j's:

$$Q_j = qQ_{j-1} + pqQ_{j-2} + p^2q^2Q_{j-4} - p^3q^3Q_{j-6}, \qquad j > 5. \quad (18)$$

The generating function is

$$Q(s) = \frac{1 + ps + p^2s^2 + p^2qs^3 - p^3qs^4 - p^3q^2s^5}{1 - qs - pqs^2 - p^2q^2s^4 + p^3q^3s^6}, \quad (19)$$

and the expected value of y is

$$E(y) = \frac{1 + p + p^2 + p^2q^3}{p^3(1 + q + q^3)}. \quad (20)$$

4.33 *The First Occurrence of Four Successes in Five Consecutive Trials*

Here the $4\cdot5/2 = 10$ initial Q_j's are:

$$\begin{aligned}
Q_0 &= Q_1 = Q_2 = Q_3 = 1, \qquad Q_4 = 1 - p^4,\\
Q_5 &= Q_4 - 4p^4q,\\
Q_6 &= Q_5 - 4p^4q^2,\\
Q_7 &= Q_6 - 4p^4q^3 - 3p^5q^2,\\
Q_8 &= Q_7 - 4p^4q^4 - 7p^5q^3 - 2p^6q^2,\\
Q_9 &= Q_8 - 4p^4q^5 - 11p^5q^4 - 9p^6q^3 - p^7q^2.
\end{aligned}$$

For $j > 9$ the following recursion equation holds:

$$\begin{aligned}
Q_j = qQ_{j-1} + pqQ_{j-2} + p^2q^2Q_{j-4} + 2p^3q^2Q_{j-5}&\\
-p^4q^3Q_{j-7} - p^6q^4Q_{j-10}.&
\end{aligned} \quad (21)$$

The generating function of Q_j is

$$Q(s) = \frac{1 + ps + p^2s^2 + p^2s^3 + 2p^3qs^4 - p^4qs^5 - p^4q^2s^6 - p^6q^2s^8 - p^6q^3s^9}{1 - qs - pqs^2 - p^2q^2s^4 - 2p^3q^2s^5 + p^4q^3s^7 + p^6q^4s^{10}}. \quad (22)$$

Then

$$E(y) = \frac{1 + p + 2p^2 + 2p^3q - p^4q - p^4q^2 - p^6q^2 - p^6q^3}{p^3(1 + q - 2q^2 + pq^3 + p^3q^4)}. \quad (23)$$

4.4 *Properties of Two-Sided Zone Tests*

The results presented in Sections 4.2 and 4.3 are applicable to the study of the statistical properties of one-sided zone tests for runs of

points in the zone above an upper limit line at $\bar{X}_0' + L_k(\sigma'/\sqrt{n})$. Generally, we also test for the same types of runs below a lower limit line at $\bar{X}_0' - L_k(\sigma'/\sqrt{n})$, in which case the test is a two-sided zone test and each point falls into one of three mutually exclusive zones.

Let A denote the zone above the upper limit, B denote the zone between the two limits, and C denote the zone below the lower limit. Consider an infinite sequence of independent trials having possible outcomes a, b, and c with fixed probabilities a, b, and c. When the outcome of the jth trial completes a pattern of outcomes describing an event ϵ we say that ϵ occurs on the jth trial. Event ϵ is defined by a set of outcome patterns and a counting, or testing, rule. If when ϵ occurs on the jth trial we treat trial $j + 1$ as though it were the first trial, ignoring the results of the first j trials, then ϵ is a recurrent event.[8]

Let

u_j = Probability that ϵ occurs on the jth trial,

f_j = " " " " for the first time on the jth trial,

Q_j = Probability that ϵ does not occur in the first j trials.

Denote the generating functions of u_j, f_j, and Q_j by $U(s)$, $F(s)$, and $Q(s)$, respectively.

The following equation can be used to determine the Q's in terms of the f's:

$$Q(s) = \frac{1 - F(s)}{1 - s} \qquad -1 < s < 1. \tag{24}$$

If ϵ is a *recurrent* event the following equation holds [Reference 8, p. 243]:

$$u_j = f_j + f_{j-1}u_1 + f_{j-2}u_2 + \cdots + f_1 u_{j-1}. \tag{25}$$

Equation (25) leads to the following identity (setting $f_0 = 0$, $u_0 = 1$):

$$U(s) = \frac{1}{1 - F(s)}. \tag{26}$$

From (24) and (26) we have

$$(1 - s)\,U(s) = \frac{1}{Q(s)}, \qquad -1 < s < 1, \tag{27}$$

for recurrent event ϵ. We shall consider only recurrent events which have finite recurrence times; in these cases $F(1) = f_1 + f_2 + \cdots = 1$, and $U(1)$ is infinite. The limit of $(1 - s)U(s)$ as s approaches unity from below is (using L'Hospital's Rule):

$$\lim_{s\to 1} (1 - s)\,U(s) = \frac{1}{F'(1)} = \frac{1}{Q(1)} = \frac{1}{E(y)}, \tag{28}$$

where y denotes the number of the trial of the first occurrence of ϵ, and $E(y)$ denotes its expected value. $E(y)$ is also the average recurrence time (average number of trials between consecutive occurrences) of recurrent event ϵ.

Consider recurrent events ϵ_1, ϵ_2, and ϵ_{12}, defined, respectively, by the sets of outcome patterns α, β, and α or β and a counting rule that requires counting to start from scratch on trial $j (j > 1)$ if and only if the event under consideration occurs on trial $j - 1$. Assume that ϵ_1 and ϵ_2 are mutually exclusive — that is, they cannot both occur on the same trial.

For an example, let the single pattern $a\ c\ a$ define the set α and the pattern $c\ a\ c$ define the set β — then the set α or β has the two patterns $a\ c\ a$ and $c\ a\ c$. Consider an outcome sequence:

trial number:	1	2	3	4	5	6	7	8	9
trial outcome:	*a*	*c*	*a*	*c*	*a*	*c*	*a*	*b*	*a*

The event ϵ_1 occurs on trials 3 and 7; the event ϵ_2 occurs on trial 4; and the event ϵ_{12} occurs on trials 3, 6, and 9.

Let $E_1(y)$, $E_2(y)$, and $E_{12}(y)$ denote the average recurrence times of ϵ_1, ϵ_2, and ϵ_{12}, respectively. Under what conditions can we determine $E_{12}(y)$ from known values of $E_1(y)$ and $E_2(y)$?

Consider events ϵ_1'' and ϵ_2'' defined by outcome patterns α and β, respectively, and a counting rule that requires counting to start from scratch on trial j if and only if *either* ϵ_1'' *or* ϵ_2'' occurs on trial $j - 1$. Events ϵ_1'' and ϵ_2'' differ from ϵ_1 and ϵ_2 only in counting rules. In the example previously considered, we see that ϵ_1'' occurred on trials 3 and 9, and ϵ_2'' occurred on trial 6. Either ϵ_1'' or ϵ_2'' (but not both) occurs on every trial on which ϵ_{12} occurs; this leads to the equation

$$u_{12,j} = u_{1,j}'' + u_{2,j}'', \tag{29}$$

where $u_{12,j}$, $u_{1,j}''$, and $u_{2,j}''$ denote, respectively, the probabilities that ϵ_{12}, ϵ_1'', and ϵ_2'' occur on trial j.

Multiplying (29) through by s^j and summing over j from one to infinity, we obtain an equation relating the generating functions of the probabilities in (29):

$$U_{12}(s) = U_1''(s) + U_2''(s) - 1. \tag{30}$$

The constant appears because $u_0 = 1$ in all cases.

Events ϵ_1'' and ϵ_2'' are recurrent events, and equations (25) through (28) can be used to determine their mean recurrence times $E_1''(y)$ and $E_2''(y)$. (The fact that (25) applies allows us to call ϵ_1'' and ϵ_2'' recurrent events).

If we multiply through (30) by $(1 - s)$ and take the limit of each side as s approaches unity (see (28)), we obtain

$$\frac{1}{E_{12}(y)} = \frac{1}{E_1''(y)} + \frac{1}{E_2''(y)}. \tag{31}$$

In any sequence of trial outcomes, if a pattern in α occurs for the first time on trial j, then ϵ_1 occurs for the first time on trial j, and ϵ_1'' occurs for the first time either on trial j or on a *later* trial; ϵ_1'' will occur for the first time on a later trial if ϵ_2'' occurred while this first pattern in α was being formed. Thus we have

$$E_1(y) \leqq E_1''(y) \tag{32}$$

where the equality sign holds if and only if no pattern in β overlaps a pattern in α. A pattern in β overlaps a pattern in α if the terminating outcomes of the former correspond to the beginning outcomes of the latter. Thus outcome pattern $c\ a\ c$ overlaps $a\ c\ a$ because the terminating outcomes $a\ c$ of the former correspond to the beginning outcomes $a\ c$ of the latter. If no pattern in β overlaps a pattern in α then the occurrence ϵ_2'' does not "cancel out" the beginning of any patterns in α, and therefore ϵ_1 and ϵ_1'' always occur on the same trials.

From (31) and (32) we have

$$\frac{1}{E_{12}(y)} \leqq \frac{1}{E_1(y)} + \frac{1}{E_2(y)}, \tag{33}$$

where the equality sign holds if and only if ϵ_1 and ϵ_2 are defined by non-overlapping patterns, in which case we shall say that ϵ_1 and ϵ_2 are non-overlapping events. From our example it is clear that mutually exclusive events are not necessarily non-overlapping.

We can use (33) to find $E(y)$ for two-sided tests in terms of the $E^*(y)$'s of the component one-sided tests. Note that a given change Δ looks like a $-\Delta$ to one of the component tests. Then

$$\frac{1}{E(y;\Delta)} \leqq \frac{1}{E^*(y;\Delta)} + \frac{1}{E^*(y;-\Delta)}. \tag{34}$$

For $\Delta = 0$, $E(y;0) \leqq (E^*(y;0)/2)$. The equality sign holds in (34) for $T_k(L_k)$ and $T_{1k}(L_1, L_k)$. For $T_{k'}(L_k)$ and $T_{1k'}(L_1, L_k)$, (34) defines lower bounds which are very close approximations to $E(y)$. For $T_{2'}$ equation (34) leads to a lower bound of 510.6, which compares with the true value $E(y;0) = 510.7$. The degree to which the approximation approaches the true value depends on the probability of overlap, which in cases we consider is very small; for this reason we can consider (34) to be an approximation rather than a lower bound.

4.5 *Properties of Tests Combining $T_1(L_1)$ with One Other Test Whose Properties Are Known*

Consider any test T for which we partition the control chart into mutually exclusive zones $A, B, C, \cdots, R$. The possible outcomes of the first j trials can be enumerated by an ordered expansion of

$$(a + b + c + \cdots + r)^j.$$

With the letters denoting the probabilities of points falling into the various zones ($a + b + c + \cdots + r = 1$), we pick out all of those terms corresponding to outcomes in which the event ϵ does not occur, and denote their sum by

$$Q_{\iota,j} = g_j(a, b, c, \cdots r). \tag{35}$$

Clearly $Q_{\iota,j}$, or g_j, is the sum of a series of terms such as $a^2bc^3 \cdots$, representing the probabilities of particular outcomes.

If we wish to find the probability $Q_{1\iota,j}$ of no occurrences of event ϵ *and* no occurrence of a point falling in zones A or B, say, we simply eliminate from g_j those terms in which either a or b occurs. We can do this by substituting zeros for a and b wherever they occur in g_j : $Q_{1\iota,j} = g_j(0, 0, c, d, \cdots, r)$. By multiplying and dividing each remaining term in $g_j(0, 0, c, d, \cdots, r)$ by $(1 - a - b)^j$, we derive an alternative expression:

$$Q_{1\iota,j} = g_j\left(0, 0, \frac{c}{1 - a - b}, \frac{d}{1 - a - b}, \cdots, \frac{r}{1 - a - b}\right)(1 - a - b)^j, \tag{36}$$

showing that the conditional probability of no ϵ given no points in A or B uses the same function required for $Q_{\iota,j}$. This enables us to write the generating function of $Q_{1\iota,j}$ as

$$Q_{1\iota}(s) = h\left(0, 0, \frac{c}{1 - a - b}, \frac{d}{1 - a - b}, \cdots, \frac{r}{1 - a - b}; (1 - a - b)s\right), \tag{37}$$

where $h(a, b, c, \cdots, r; s)$, defined for $a + b + c + \cdots + r = 1$, is the generating function of $Q_{\iota,j}$.

The principles are best illustrated by an example. Consider the problem of finding $E_{1k}^*(y)$, the expected number of the trial of the first occurrence of k consecutive points above $\bar{X}_0' + L_k(\sigma'/\sqrt{n})$ *or* a single point

above $\bar{X}_0' + L_1(\sigma'/\sqrt{n})$, where $L_k < L_1$. We use an asterisk to denote the fact that a function applies to a one-sided test. We let a be the probability of a point falling above both limits, b be the probability of falling between the two limits, and c be the probability of falling below both limits. Then we substitute $a + b$ for p, and c for q in (7) to obtain

$$Q_k^*(s) = \frac{1 - (a + b)^k s^k}{1 - s + c(a + b)^k s^{k+1}}. \tag{39}$$

Following (37), we find $Q_{1k}^*(s)$ by substituting 0 for a, $b/(1 - a)$ for b, $c/(1 - a)$ for c, and $(1 - a)s = (b + c)s$ for s in (39),

$$Q_{1k}^*(s) = \frac{1 - b^k s^k}{1 - (b + c)s + cb^k s^{k+1}}. \tag{40}$$

We set $s = 1$ in (40) to obtain

$$E_{1k}^*(s) = \frac{1 - b^k}{1 - (b + c) + cb^k}. \tag{41}$$

The properties of any test combining $T_1(L_1)$ with one other test whose properties are known can be determined in a similar way.

4.6 *Limits of $E(y)$ in Compound Tests*

A development similar to that in Section 4.4 will show, for example, that

$$\frac{1}{E_{123}(y)} \leqq \frac{1}{E_1(y)} + \frac{1}{E_2(y)} + \frac{1}{E_3(y)}, \tag{42}$$

where $E_{123}(y)$ pertains to recurrent event ϵ_{123}, whose set of outcome patterns is composed of those of recurrent events ϵ_1, ϵ_2, and ϵ_3.

It can also be shown that

$$E_{123}(y) \leqq E_{12}(y) \leqq E_1(y) \tag{43}$$

for example. Clearly we cannot increase the recurrence time of an event by increasing the different outcome patterns which define the event.

4.7 *Translating Limits to Obtain a Selected Probability of Type 1 Error*

By supplementing T_1 with other tests, we increase the probability of Type 1 errors. We can adjust the probability of Type 1 errors to any desired level by resetting the zone limits. With more than one set of limit lines, we have some freedom in setting the limits. A procedure that has the important attribute of simplicity translates all of the limit lines away from the central line $\bar{X}_0'$ by the same amount. The properties of the resultant test can be derived directly from the properties of the original test.

We first determine the properties of one-sided tests whose limits are translated; from these results we determine the properties of the corresponding two-sided test.

If $Q_t^*(L_{k_1}, L_{k_2}; \Delta; s)$ is the generating function of Q_j for test T^* with limits at $\bar{X}_0' + L_{k_1}(\sigma'/\sqrt{n})$ and $\bar{X}_0' + L_{k_2}(\sigma'/\sqrt{n})$ and with the process average $\bar{X}' = \bar{X}_0' + \Delta$, then (neglecting points from State 1):

$$Q_t^*(L_{k_1}, L_{k_2}; \Delta; s) \equiv Q_t^*\left(L_{k_1} + h, L_{k_2} + h; \Delta - h\frac{\sigma'}{\sqrt{n}}; s\right). \quad (44)$$

This equation says that the probabilities involved are identical if we translate the limits by a given amount or if we translate the process average in the opposite direction by the same amount. The truth of this stems from the fact that the probabilities depend on the position of the process average $\bar{X}'$ *relative* to the zone limits.

If we wish to set the limits so that the probability of a Type 1 error is $\frac{1}{500}$, say, for a two-sided test, we can proceed as follows:

(1) draw the curve of $E(y)$ versus Δ for the corresponding one-sided test (the abscissa is assumed to be in units of $\sigma'/\sqrt{n}$),

(2) translate this curve to the right (or left) until $E(y; 0) = 1000$,

(3) measure the amount h of the translation, and translate the zone limits away from (or toward) $\bar{X}_0'$ by an amount $h(\sigma'/\sqrt{n})$ (control chart units).

The translated $E(y)$ versus Δ curve represents the new one-sided test. The curve for the corresponding two-sided test can be derived using (34); it will have a value $E(y; 0) = 500$.

4.8 *Monte Carlo Techniques to Determine the Properties of Zone Tests*

We can determine approximately the properties of zone tests by using Monte Carlo techniques on modern high-speed computers. First we generate a random series of numbers with a known distribution. Then, using the appropriate correspondence between limits within the distribution and zone limits, we translate the random numbers into a random sequence of zone designations, which we test for occurrences of the events in question. We keep score of the number of points until the event finally occurs. We then start counting again as though the sequence were just starting. By running through a great many cycles, we obtain an approximation to the distribution of the cycle length y, and an approximation to $E(y)$ for the particular value of Δ that applies to the limits we used. We repeat the process with different limits for different values of Δ.

Within the limitations of the computer, this technique can be used for any zone test. We used it to approximate the properties of $T_{12'4'8}$ for $\Delta = 0, \sigma'/\sqrt{n}, 2(\sigma'/\sqrt{n})$, and $3(\sigma'/\sqrt{n})$.

V. CONCLUSIONS

If we supplement the standard control chart test with another zone test, we increase its sensitivity to process changes at the cost of more

frequent errors of Type 1 and a more complicated testing procedure; see Figs. 5 and 6. We can restore the original probability of Type 1 errors by changing the zone limits; in the following discussion we shall assume that this has been done, thereby simplifying the comparison of various tests. We shall say that a zone test T_t of the type we are concerned with is better than T_1 for a particular value of Δ if $E_t(y) < E_1(y)$ for that Δ.

In general, the curve of $E(y)$ versus Δ for a test T_t is below the corresponding curve for T_1 for Δ in a range $0 < \Delta < \Delta_t$, and above for $\Delta > \Delta_t$. The crossover point Δ_t in the cases we considered varied from 1.7 $(\sigma'/\sqrt{n})$ for $T_8(0.065)$ (Fig. 7) to over 3.5 $(\sigma'/\sqrt{n})$ for $T_{12'}(3.13, 2.13)$ (Fig. 9).

Consider a test T_{1t} that combines T_1 and T_t and has its zone limits set so that its probability of Type 1 errors is the same as for T_1 and for T_t. In the cases we have considered (see Figs. 8 and 9) T_{1t} essentially effects a compromise between T_1 and T_t — for small changes it is better than T_1 but not as good as T_t; for large changes it is better than T_t but not as good as T_1; for Δ near Δ_t it is better than T_1 and better than T_t.

In the cases we have considered, tests $T_{k'}(L_{k'})$ appear to be slightly better than tests $T_k(L_k)$ for small changes.

The reason that zone test T_t is better than T_1 for small changes seems to be due to the fact that it bases its decisions on a history of k consecutive points; in effect, it makes some use of a sample size kn rather than n. The cost of the increased effective sample size is paid during the first $k - 1$ points in State 2, where T_1 has a higher probability than T_t of detecting a change. The probability that a point falls outside of a 3σ limit remains fixed from sample to sample, and after the initial $k - 1$ points in State 2, this probability is less than the probability that T_t will detect a run. Large changes are likely to be detected by T_1 before T_t becomes effective; but when changes are small the corresponding values of $E(y)$ are large, and we can expect T_t to detect the change before T_1 does (see Fig. 7).

We have assumed that sample averages $\bar{X}_n$ are plotted on the control chart. In light of the above discussion the possibility of pooling data from k consecutive samples and plotting a statistic based on the kn measurements involved appears promising.

A preliminary study of zone tests on charts of moving averages of k consecutive equal-sized samples has been made. The statistic (or point)

$$Y_{kn,j} = (\bar{X}_{n,j} + \bar{X}_{n,j-1} + \cdots + \bar{X}_{n,j-k+1})/k$$

can easily be determined graphically in many cases. For example, the point $Y_{2n,j}$ is halfway between points $\bar{X}_{n,j-1}$ and $\bar{X}_{n,j}$ on the straight

line connecting them — vertical rulings on cross-section paper ordinarily used will spot points exactly. Points $Y_{4n,j}$ can be similarly derived from points $Y_{2n,j-2}$ and $Y_{2n,j}$. Fig. 12 shows curves of $E(y)$ versus Δ for T_1 used on points $Y_{kn,j}(k = 1, 2, 4)$; limit lines were assumed to be at $\bar{X}_0' \pm 3(\sigma'/\sqrt{kn})$. The curve for $k = 1$ is, of course, the curve T_1 of earlier figures; the curve for $k = 2$ was derived using tables of the bivariate normal distribution; the curve for $k = 4$ is an approximation based primarily on the results of a study making use of Monte Carlo

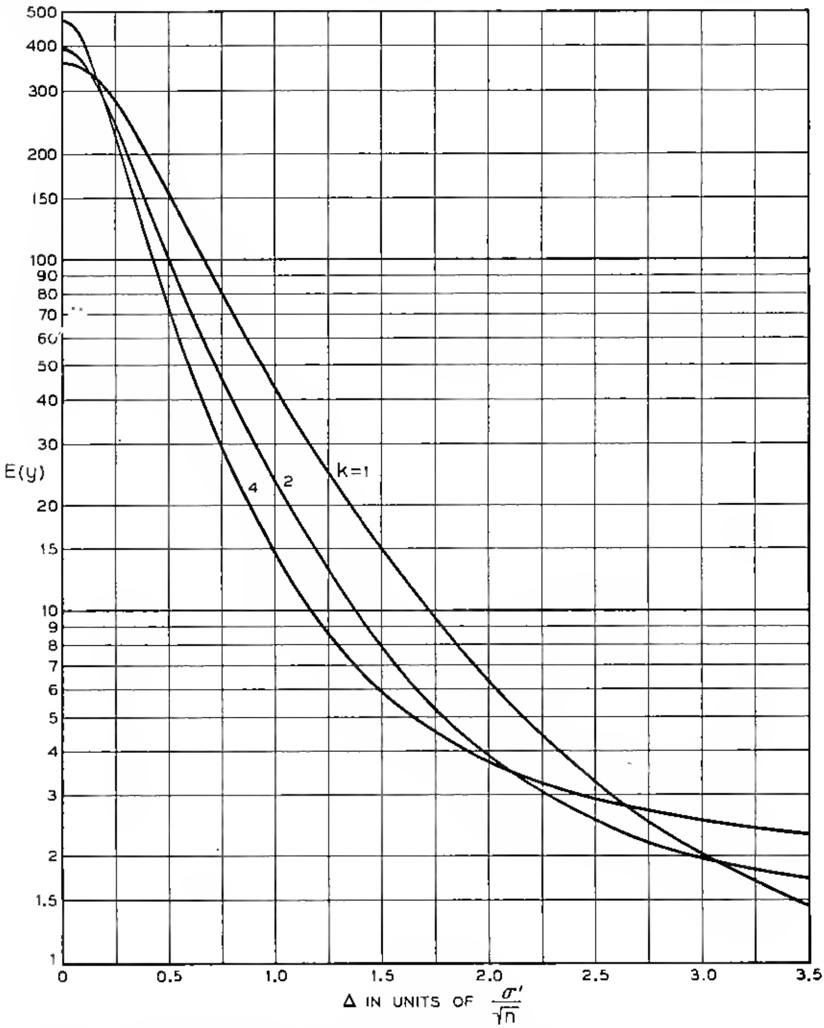

FIG. 12 — $E(y)$ versus Δ for T_1 applied to moving averages of $k(k = 1, 2, 4)$ consecutive sample averages.

techniques. We cover the possibility that the first point in State 2 will fall outside of its control limits by assuming the existence of prior points in State 1; all three curves approach $E(y) = 1$ as Δ approaches infinity.

In comparing Fig. 12 with Figs. 7, 8, and 9, it appears that T_1 used on moving averages provides an effective test for detecting shifts in process averages. Further study is required to determine the effectiveness of other run tests and of combinations of run tests applied to various moving averages.

In summary, it is possible to devise zone tests which — within the constraints of our model:

(1) indicate changes in process averages when none has occurred with the same average frequency as the standard control chart test T_1,

(2) detect small changes in process average — up to $1.3\sigma'$, say, for $n = 5$ — sooner on the average than T_1, and

(3) detect larger changes inappreciably later on the average than T_1. Such tests require an appropriate setting of zone limits — generally at non-integral multiples of $\sigma'/\sqrt{n}$. If run tests are used to supplement T_1 without a compensating setting of zone limits, an increased frequency of false indications of process changes results.

The standard control chart test T_1 (or $T_1(L_1)$) is slightly more effective than alternative zone tests in detecting relatively large changes; in addition, it has the important virtue of simplicity — a virtue that extends the range of economic application of T_1 into areas where alternative tests have better statistical properties. It is difficult to recommend a single alternative test to T_1 for general application, though it is clear that alternative tests may be profitably used in many applications where early detection of relatively small changes is important.

REFERENCES

1. Olmstead, P. S., How to Detect the Type of an Assignable Cause, Industrial Quality Control, **9**, pp. 32–36, 1952.
2. Shewhart, W. A., *Economic Control of Quality of Manufactured Product*, D. Van Nostrand Co., Inc., New York, N. Y., 1931.
3. American War Standard Control Chart Method of Analyzing Data, American Standards Association, Inc., New York, N. Y. Z1.2–1941.
4. Grant, E. L., *Statistical Quality Control*, McGraw-Hill, New York, 1952.
5. Duncan, A. J., The Economic Design of $\bar{X}$ Charts Used to Maintain Current Control of a Process, Journal of A.S.A., **51**, pp. 228–42, 1956.
6. Weiler, H., The Use of Runs to Control the Mean in Quality Control, Journal of the A.S.A., **48**, pp. 816–825, 1953.
7. Weiler, H., On the Most Economical Sample Size for Controlling the Mean of a Population, Annals of Math. Stat., **23**, pp. 247–54, 1952.
8. Feller, W., *An Introduction to Probability Theory and Its Applications*, Wiley and Sons, New York, 1950.
9. Kitagawa, T., and Seguchi, T., The Combined Use of Runs in Statistical Quality Control, Bulletin of Mathematical Statistics (Japan), **7**, pp. 24–45, 53–72, 1956–57.